Taxonomic & Functional Annotation

Mihai Pop
(with material from Jackie Michaelis)

What the heck is this? (Taxonomy)
What does it do? (Function)

>3492_2769

CAATTACCGCGGCTGCTGGCACGTAGTTAGCCGTGGCTTTCTGGTTAGATACCGTCAAGG

GATGAACAGTTACTCTCATCCTTGTTCTTCTCTAACAACAGAGTTTTACGATCCGAAAAC

CTTCTTCACTCACGCGGCGTTGCTCGGTCAGACTTTCGTCCATTGCCGAAGATTCCCTAC

TGCTGCCTCCCGTAGGAGTTTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCACCCTCTC

AGGTCGGCTATGCATCGTGGCCTTGGTGAGCCGTTACCTCACCAACTAGCTAATGCACCG

CGGGTCCATCCATCAGCGACACCCGAAAGCGCCTTTCAAATCAAAACCATGCGGTTTCGA

TTGTTATACGGTATTAGCACCTGTTTCCAAGTGTTACTCCCCTTCTGATGGGCAGGTTAC

CCACGTGTTACTCACCCGTTCGCCACTCCTCTTTTTCCGGTGGAGCAAGCTCCGGTAGGA

AAAAGAAGCGTTCGACTTGCATGTATTAGGCACGCCGCCAGCGTGTCTCGTAGCACGACG

Is it similar to something we know?

q_acc	subj_acc	% id	len	mis	gaps	q.s.	q.e.
3492_2769	ON729291.1	99.617	522	0	2	3	524

s.s.	s.e.	eval	bit score
545	26	0.0	952

ON729291.1 - Enterococcus faecium strain GSP2 (what it is)
16S ribosomal RNA gene (what it does)

Defining "similar"

3 main paradigms

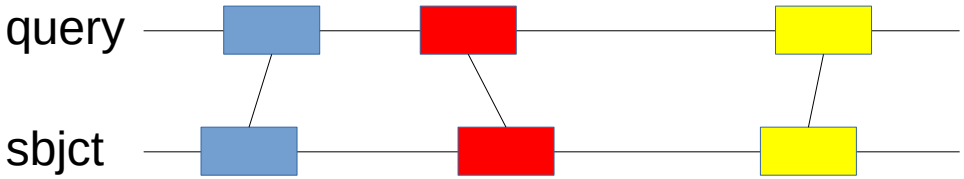
- Sequence search against a database (homology)
- Matching against machine learning models
- Inferring evolutionary processes

Database search algorithm

- Precise and slow
  - BLAST, Megan, Minimap2, etc.

```
Query   363  GTTATACGGTATTAGCACCTGTTTCCAAGTGTTACTCCCCTTCTGATGGGCAGGTTACCC  422
             ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct   161  GTTATACGGTATTAGCACCTGTTTCCAAGTGTTA-TCCCCTTCTGATGGGCAGGTTACCC  103
```

- Quick and dirty
  - Sourmash



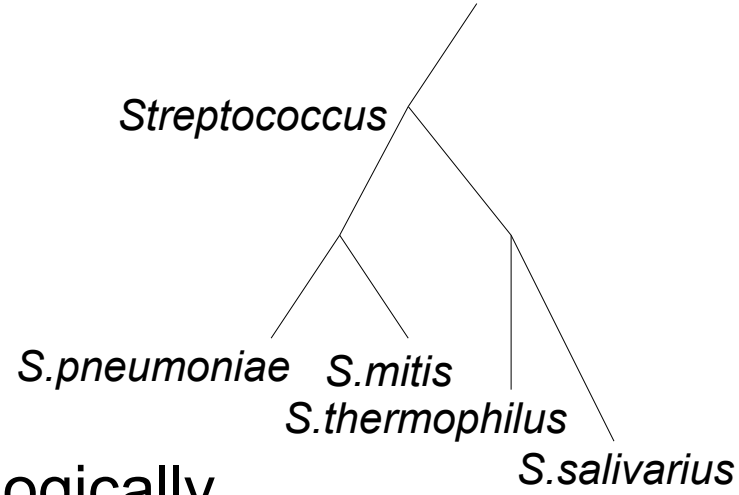
Database contents

- Genes (protein or DNA sequences)
  - 16S rRNA (highly conserved, not single copy)
  - marker genes (e.g., mOTU) (conserved, usually single copy)
  - functional genes (e.g., ARDB, HumAnn, Megan)

- Whole genomes
  - great for non-gene information
  - horizontal gene transfer may introduce errors/ambiguity

Some caveats

- Dealing with ambiguities
  - common solution:
    most recent common ancestor

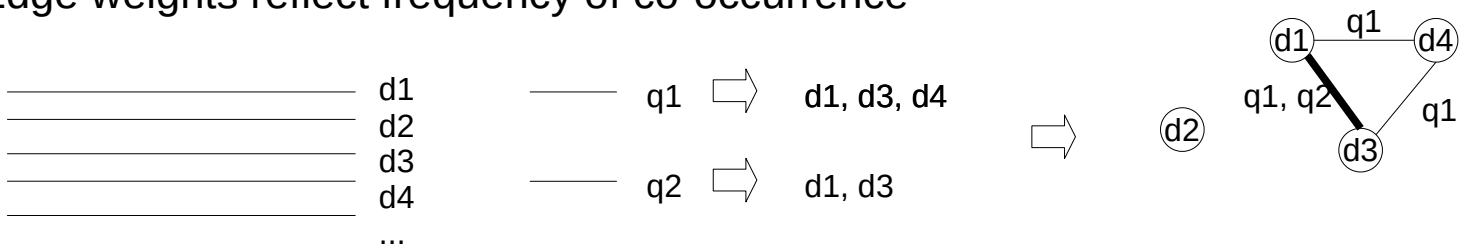


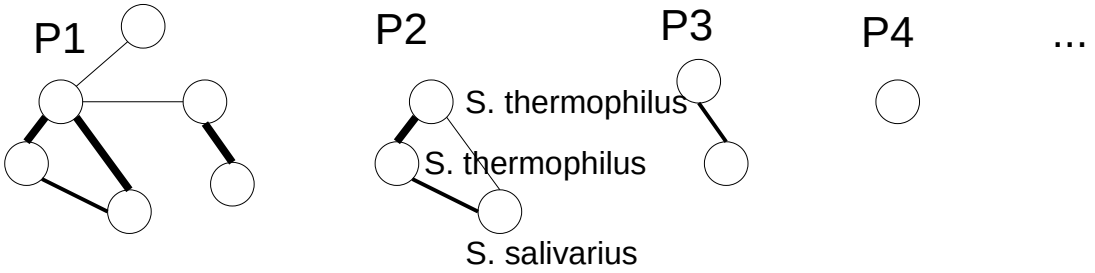
- Discriminant sequences may be biologically meaningless
  - e.g., phage protein used as discriminator in MetaPhlAn

- Database incompleteness leads to errors

ATLAS - data-driven database partitioning

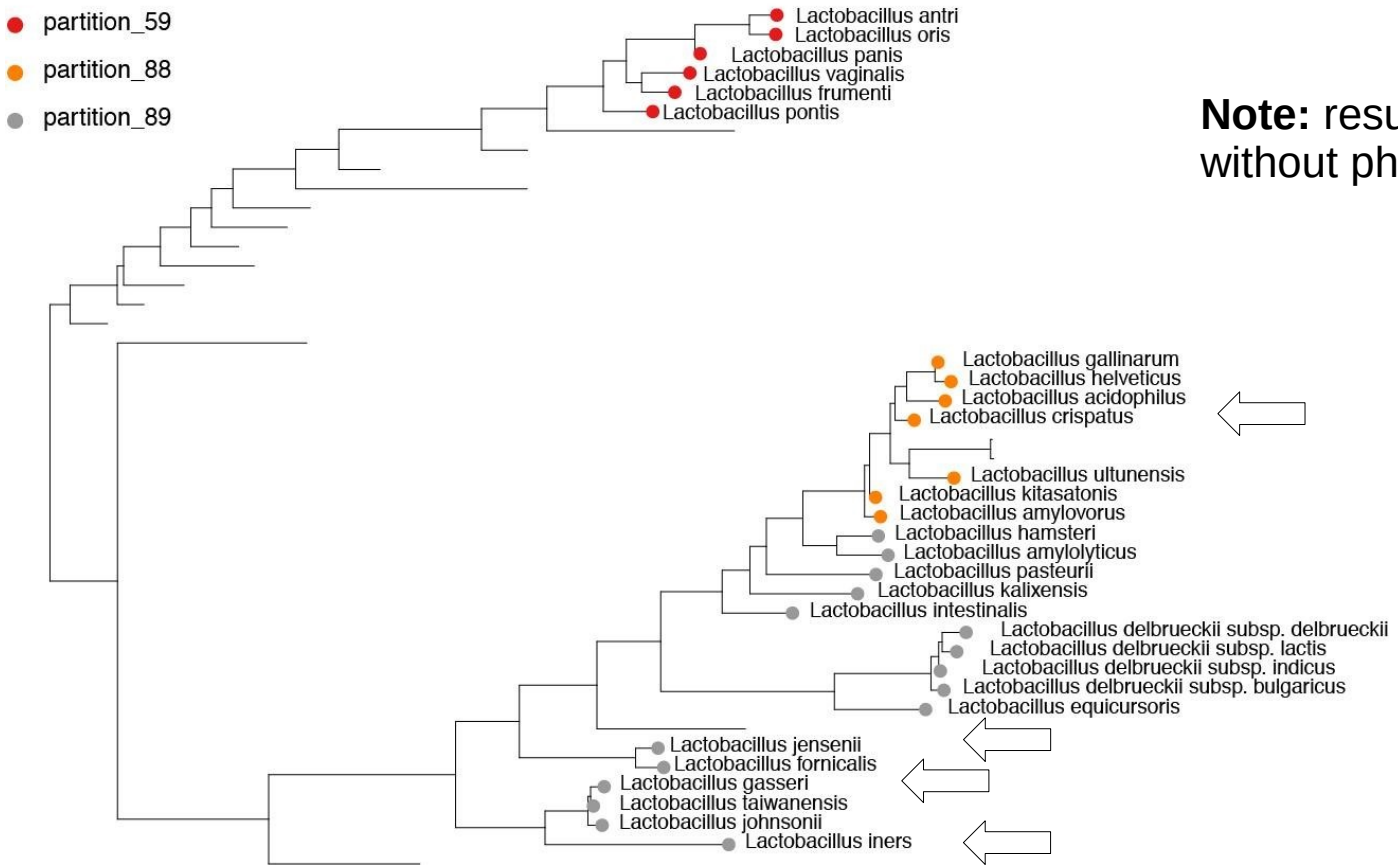
Premise: database sequences that co-occur in the outlier set are related
Edge weights reflect frequency of co-occurrence





https://github.com/marbl/ATLAS

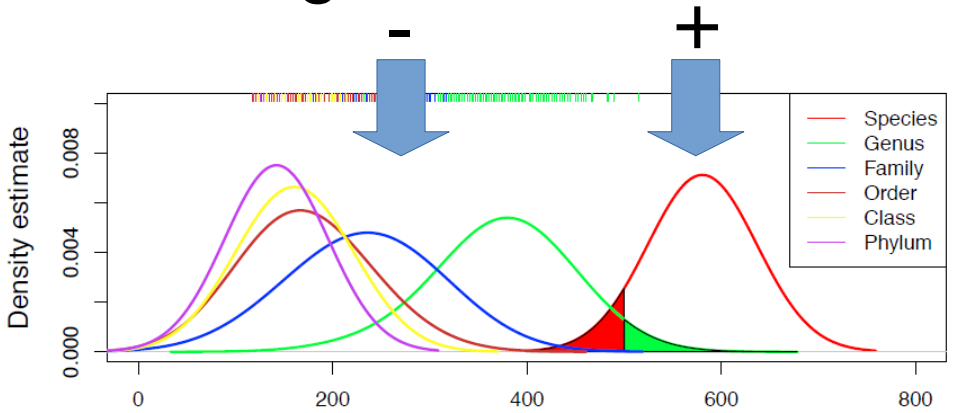
Lactobacillus in vaginal samples



Note: results obtained without phylogenetic analysis

Machine learning

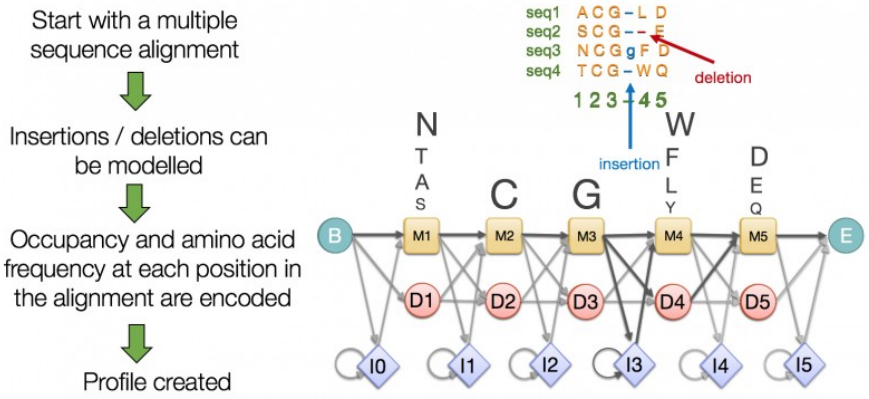
- Training data
  - "positive" examples
  - "negative" examples



- 3 Stages
  - training
  - validation (cross-validation)
  - testing

Some examples

- Alignment-based
  - Metaphyler (models bit score distribution from BLAST)
  - profile hidden Markov models (HMMs) (model columns in sequence alignment)



https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/

Some examples

- Feature-based
  - k-mer distributions (e.g., RDP classifier)

    for each 8-mer w_i in the N training sequences

$$P_i = \frac{m(w_i) + 0.5}{N + 1}$$

    for each genus

$$P(w_i|G) = \frac{m_G(w_i) + P_i}{N_G + 1}$$

    for query S

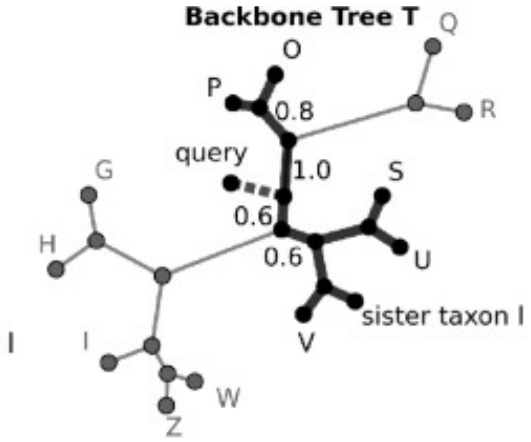
$$P(S|G) = \prod_{w_i \in S} P(w_i|G)$$

Discriminant-based

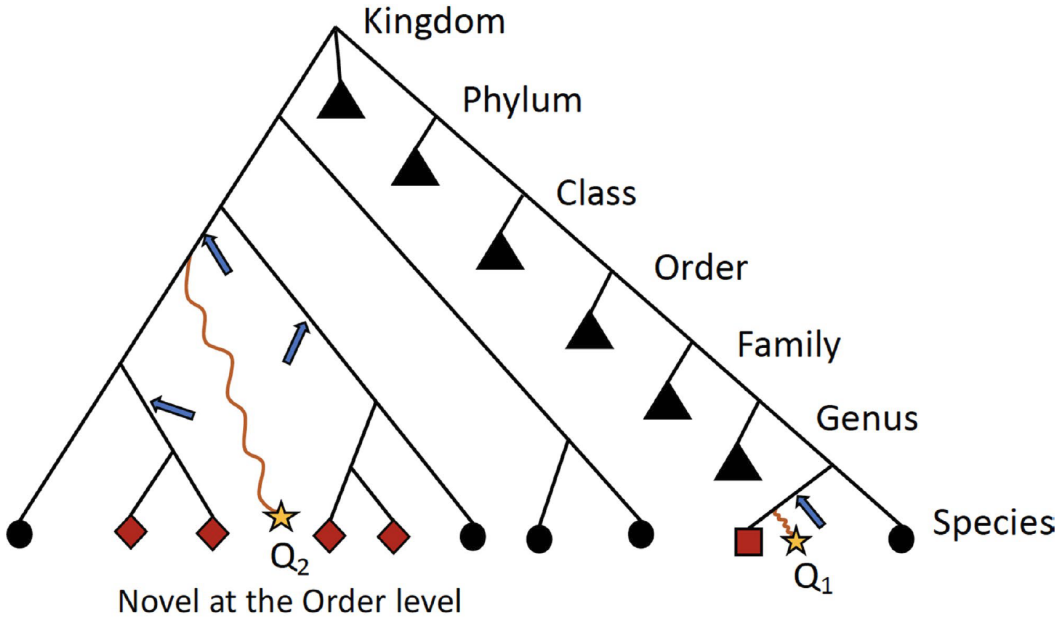
- Find sequences that discriminate between labels
  - kraken – k-mers
  - MetaPhlAn - genes

Inferring evolution: phylogenetic approaches

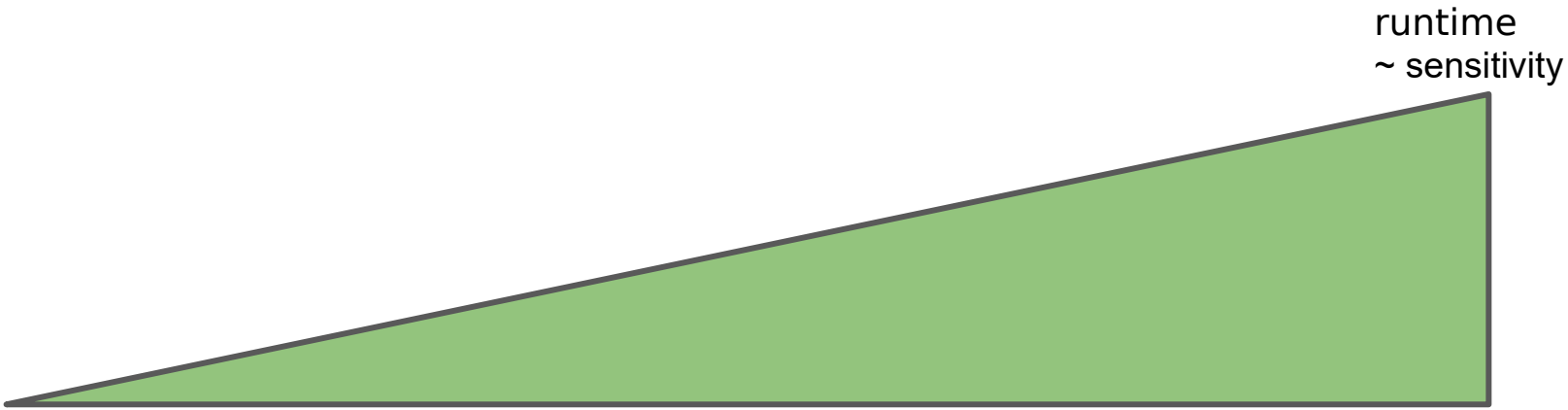
- Estimate evolutionary relationships between database sequences
- Find where the query sequence fits in the tree
- Examples: TIPP, pplacer, etc



DOI: 10.1007/978-3-030-74432-8_7



Landscape of Classification Approaches



k-mer counting
kraken

k-mer stat model
RDP

Sequence Alignment
blast
usearch
diamond

Phylogenetic
pplacer
TIPP

Assume related sequences have similar k-mer frequencies

Assume related sequences are similar

Simulate evolution

Some considerations

- Database searches
  - fast training
  - fast application
  - hard to tell if answer is incorrect
  - usually over-fit
- Machine learning
  - slow training
  - fast application
  - may over-fit
- Phylogenetic tools
  - slow training
  - slow application
  - most likely to generalize

Caveat when dealing with genes

ATTAGATGGTATTGAGACCACTGGCACAAAGATAATTTGTAC

ATTAG **ATG** GTA TTG AGA CCA CTG GCA CAA AGA **TAA** TTTGTAC

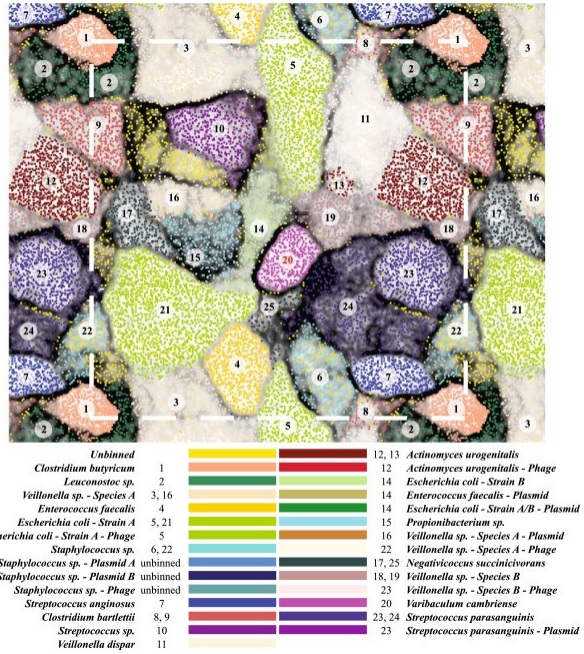
fancy machine learning algorithms

Gene or no gene?

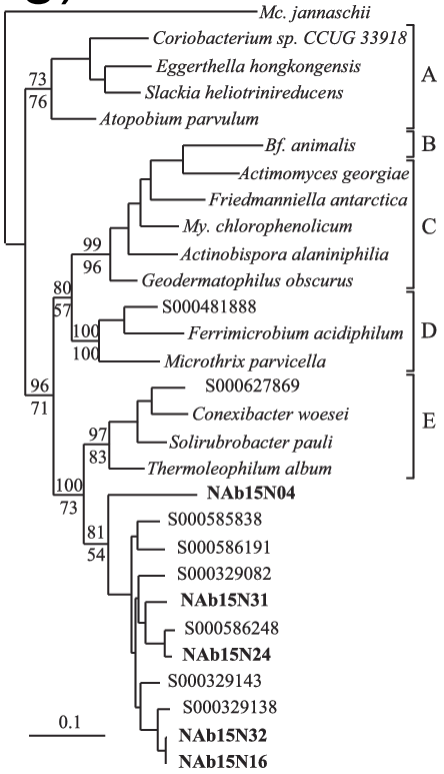
Errors in sequence or software lead to missing or truncated genes

What if you know nothing?

- Unsupervised machine learning (clustering)



DOI:10.1186/2049-2618-1-30



DOI: 10.1128/AEM.02610-06

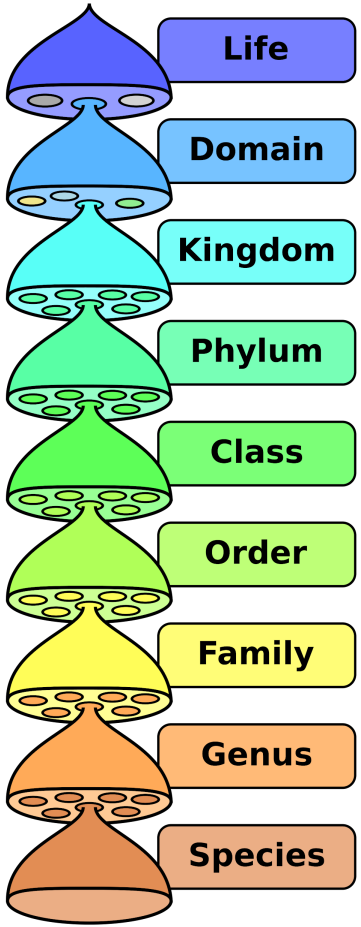
More about the labels

- What is it?
  - Taxonomy

- What does it do?
  - Gene Ontology
  - Enzyme commission numbers
  - Pathway (Kegg/BioCyc)

What is taxonomy?

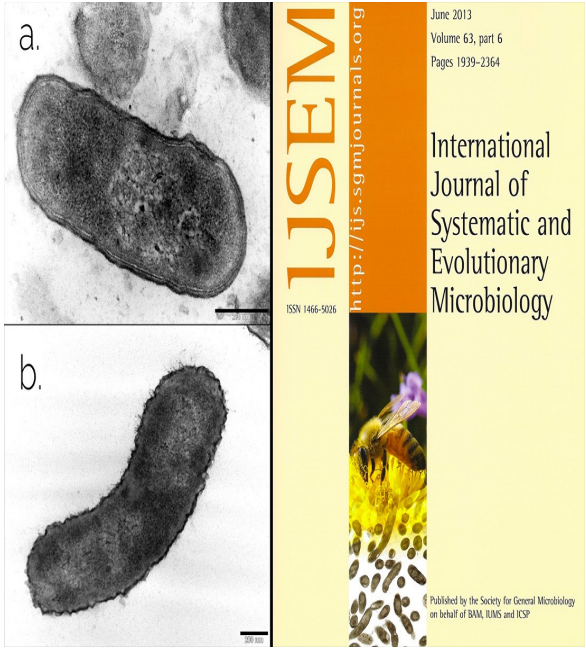
- Classification of organisms, typically arranged in hierarchical ranks (e.g. kingdom, phylum, class, order, genus, species)
- For meaningful community comparisons, taxonomic names must be consistent
- “Official” Taxonomic Names
  - Bergey’s Taxonomic Outline - manual of taxonomic names for bacteria
- "Unofficial" taxonomies
  - Many tools limit the allowable levels (e.g., no sub-divisions between genus and species)



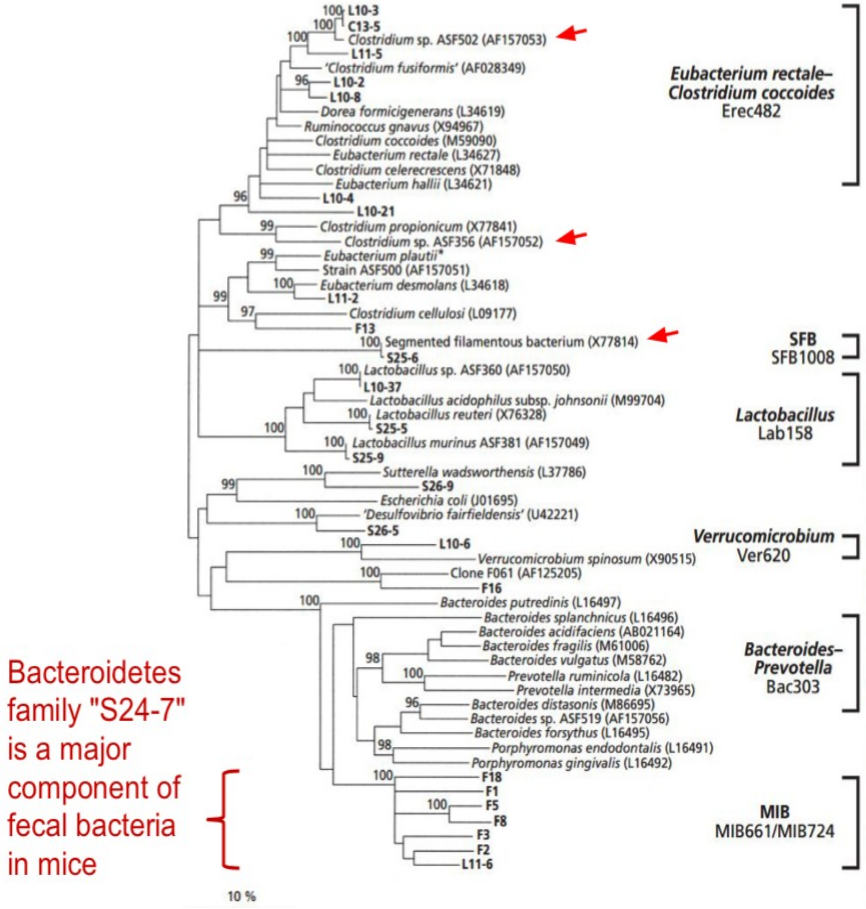
https://en.wikipedia.org/wiki/Taxonomic_rank

New bacterial species are published in IJSEM

- Species are defined by a type strain.
- To be a new species, a newly isolated strain must be sufficiently different from an existing type strain
- Requirements:
  - Genomically distinct
  - Distinguishable by phenotype
  - Deposited at two strain collections
- Genus, family, and class level taxa are also named in this system
- bacterio.net has info on named bacterial species, including links to papers and 16S sequences.
- SILVA maintains aligned 16S sequences for named species.



https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf



Bacteroidetes family "S24-7" is a major component of fecal bacteria in mice

MANY BACTERIA ARE UNNAMED

- *Clostridium* sp. ASF502 and ASF356 are part of Altered Schaedler Flora used in research for 40 years
- Segmented filamentous bacteria (SFB) were the topic of high-profile research
- Mouse intestinal bacteria (MIB) may comprise up to 80% of bacteria in feces of lab mice

SOME BACTERIA ARE NAMED INCONSISTENTLY

- *Eubacterium*, *Clostridium*, *Ruminococcus*

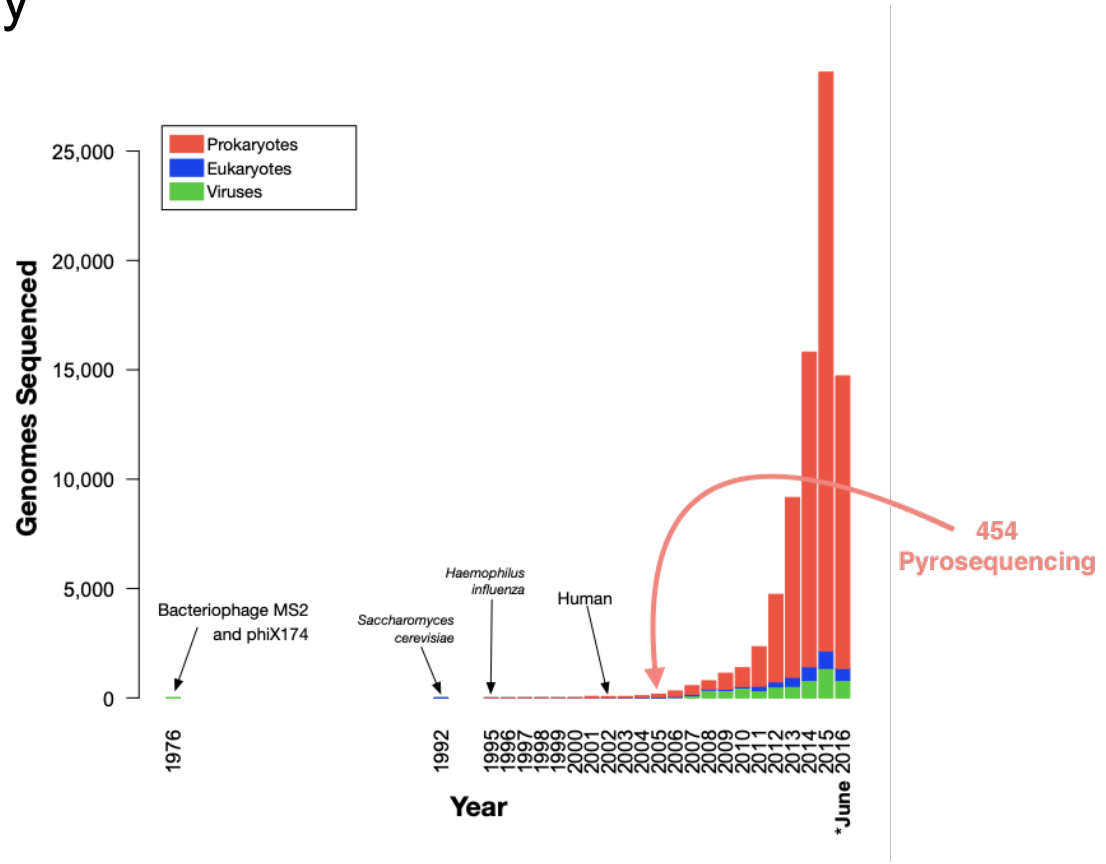
Salzman *Microbiology* **148**, 3651 (2002)

https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf

Salzman *Microbiology* **148**, 3651 (2002)

Genomes sequenced annually

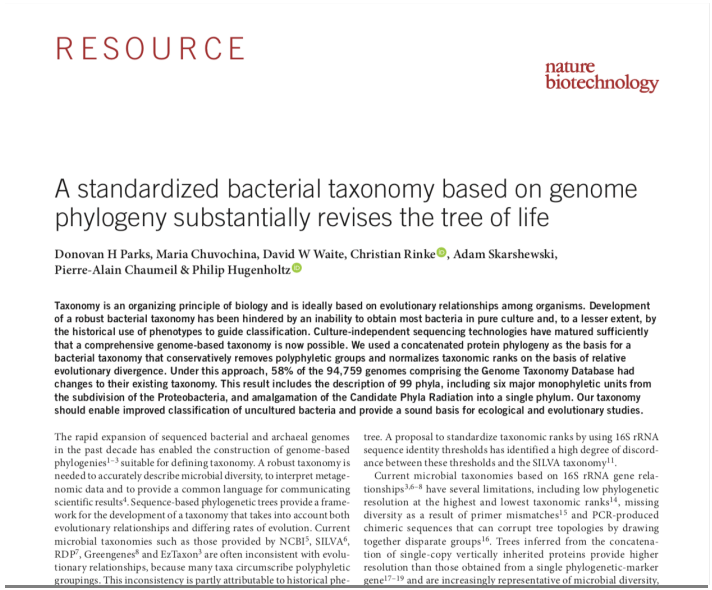
- Perhaps an overwhelming number of genomes being sequenced annually
  - Number of genomes sequenced annually doubles just about every year
- Add to this the growing *number* of “metagenome assembled genomes” (MAGs) sequenced and assembled directly from environmental samples
- Our inability to obtain most bacteria in pure culture continues to prevent the development of a robust bacterial taxonomy



Produced from NCBI GenBank files June 2016 - DJ Nasko

A new taxonomy - GTDB

- If you really want to go down the rabbit hole...



RESOURCE

nature biotechnology

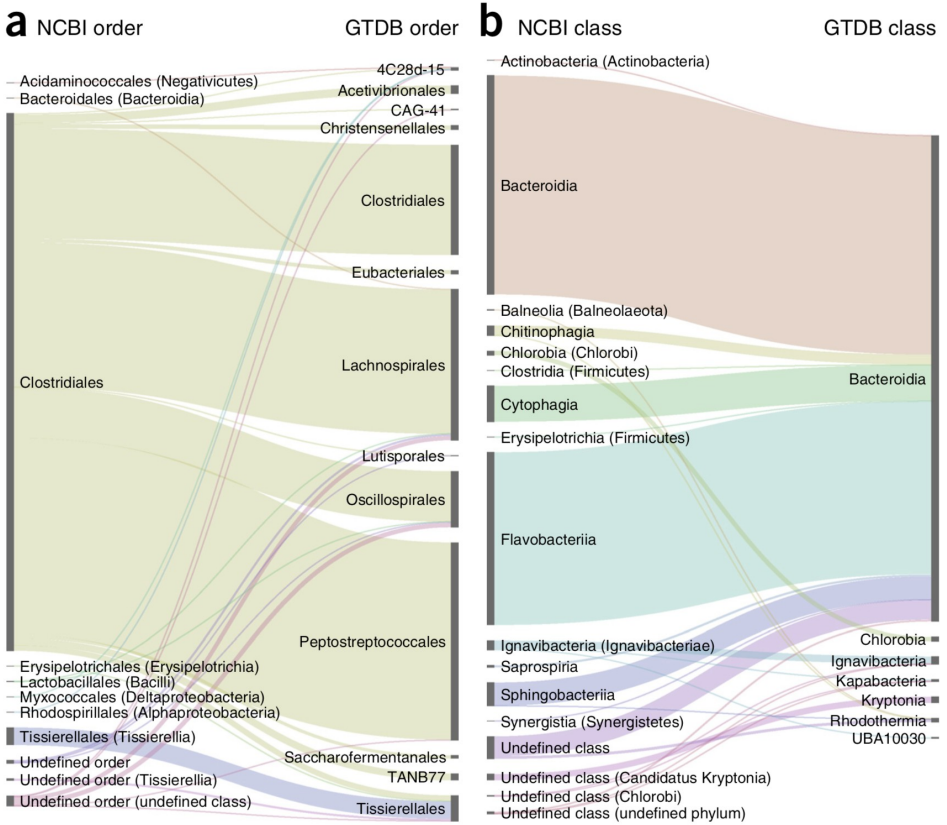
A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

The rapid expansion of sequenced bacterial and archaeal genomes in the past decade has enabled the construction of genome-based phylogenies[1–3] suitable for defining taxonomy. A robust taxonomy is needed to accurately describe microbial diversity, to interpret metagenomic data and to provide a common language for communicating scientific results[4]. Sequence-based phylogenetic trees provide a framework for the development of a taxonomy that takes into account both evolutionary relationships and differing rates of evolution. Current microbial taxonomies such as those provided by NCBI[5], SILVA[6], RDP[7], Greengenes[8] and EzTaxon[3] are often inconsistent with evolutionary relationships, because many taxa circumscribe polyphyletic groupings. This inconsistency is partly attributable to historical phenetic. A proposal to standardize taxonomic ranks by using 16S rRNA sequence identity thresholds has identified a high degree of discordance between these thresholds and the SILVA taxonomy[11].

Current microbial taxonomies based on 16S rRNA gene relationships[3,6–8] have several limitations, including low phylogenetic resolution at the highest and lowest taxonomic ranks[14], missing diversity as a result of primer mismatches[15] and PCR-produced chimeric sequences that can corrupt tree topologies by drawing together disparate groups[16]. Trees inferred from the concatenation of single-copy vertically inherited proteins provide higher resolution than those obtained from a single phylogenetic-marker gene[17–19] and are increasingly representative of microbial diversity,

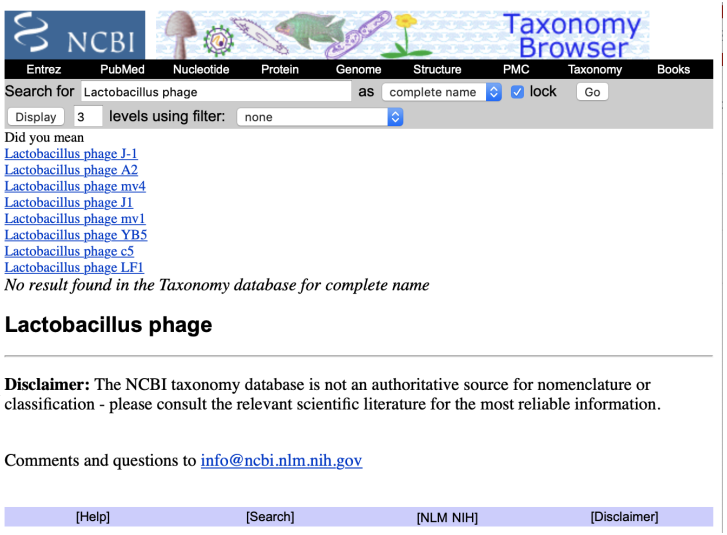


Parks, Donovan H., et al. "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life." Nature biotechnology (2018).

Common Taxonomic Systems for Bacteria

- RDP (https://rdp.cme.msu.edu/)
- Greengenes (http://greengenes.secondgenome.com/)
- SILVA (https://www.arb-silva.de/)
- NCBI (https://www.ncbi.nlm.nih.gov/taxonomy)

Note: NCBI is not an “authoritative source for nomenclature or classification”



Specialized Databases

- HOMD- Human Oral Microbiome Database (http://www.homd.org)
- OSU CORE for oral (http://microbiome.osu.edu/)
- UNITE- Fungal ITS (http://unite.ut.ee/)
- Virus-Host Database (https://www.genome.jp/virushostdb/)
- EuPathDB - Eukaryotic pathogens (https://eupathdb.org/eupathdb/)

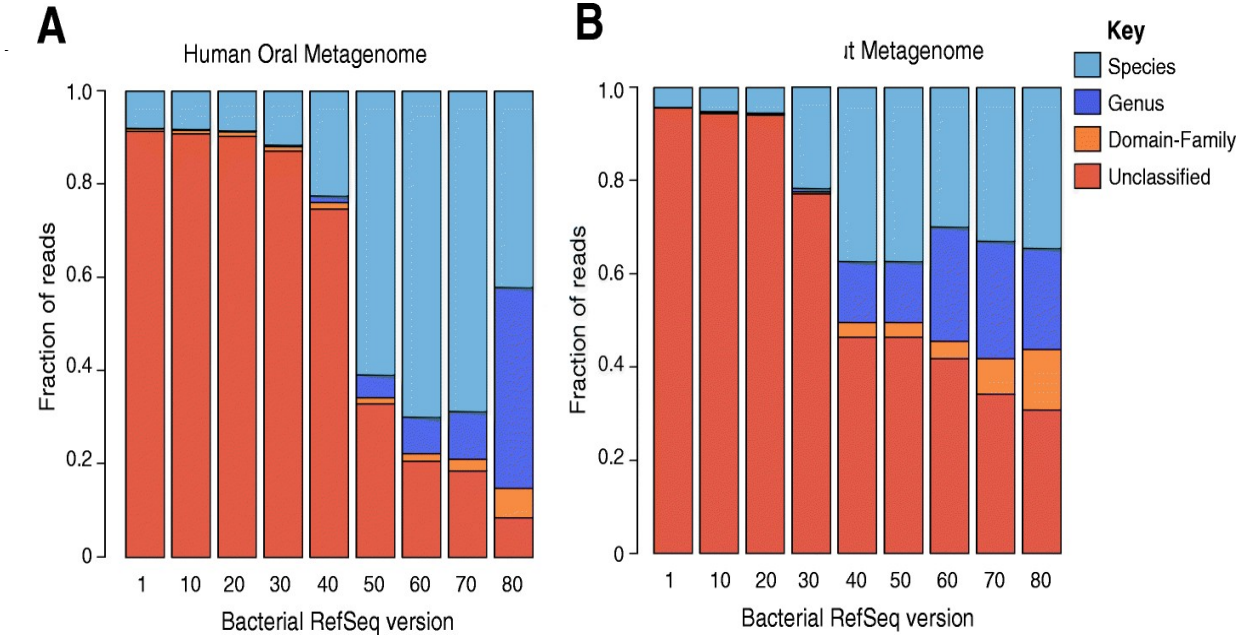
from Sue Huse

Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- e.g. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

Cross-stitch by Dr. Jennifer Glass
https://twitter.com/methanojen/status/1033749220396814336

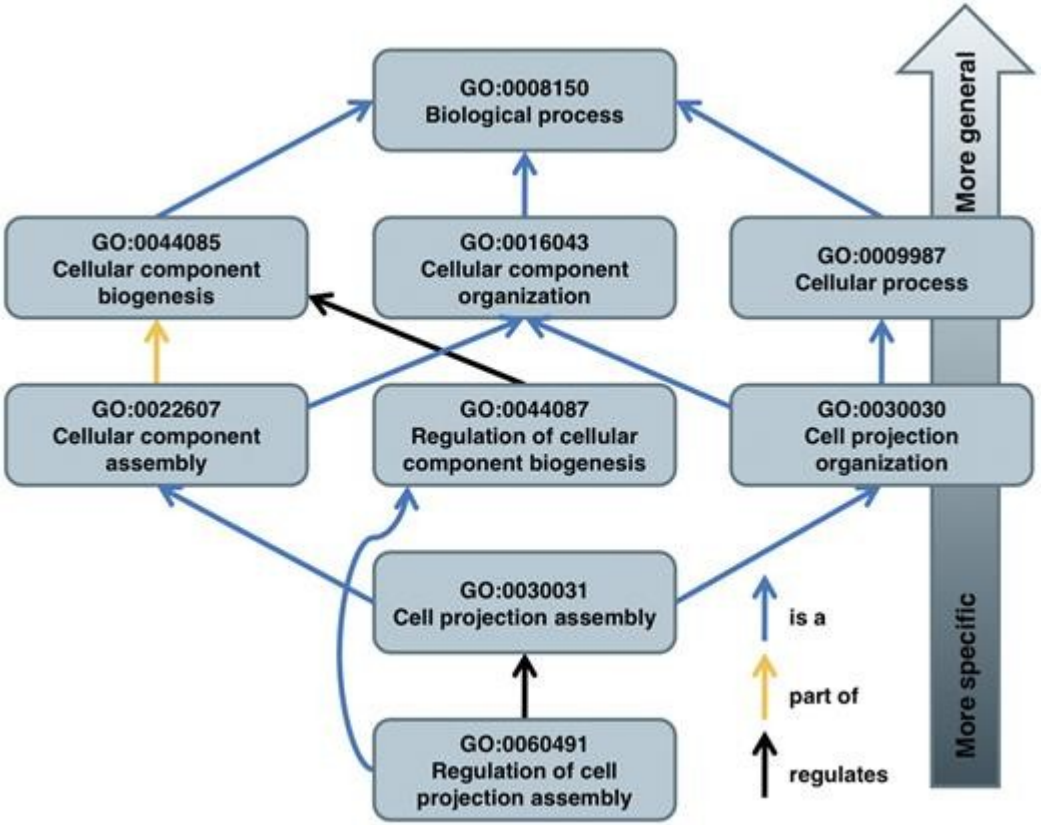
The more we know, the less we do



Nasko et al. *Genome Biology* 2018

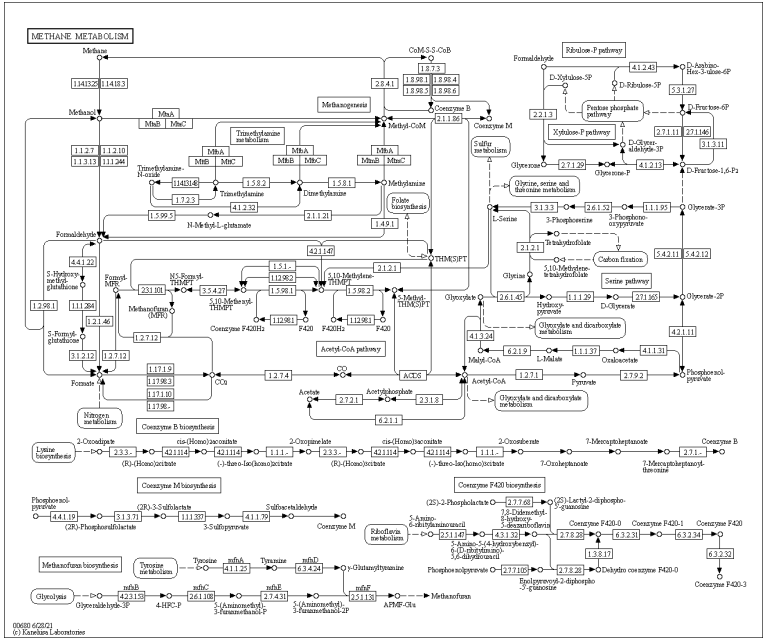
Functional labels

- Controlled vocabulary
- Hierarchical labels
- Gene Ontology

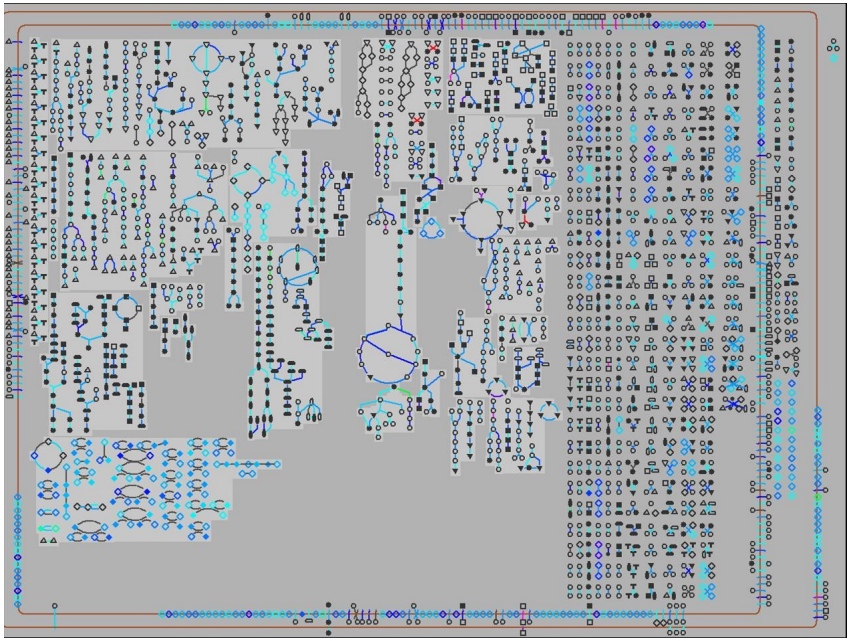


Pathways

- Basis for mechanistic models of metabolism



Kyoto Encyclopedia of Genes and Genomes (KEGG)



BioCyc/MetaCyc/EcoCyc

Enzymes

- Controlled vocabulary for describing enzymes (1950s)

EC numbers

Enzyme Commission

each enzyme has a 4-part numerical ID

more than just a name, it gives you information about what the enzyme does & how

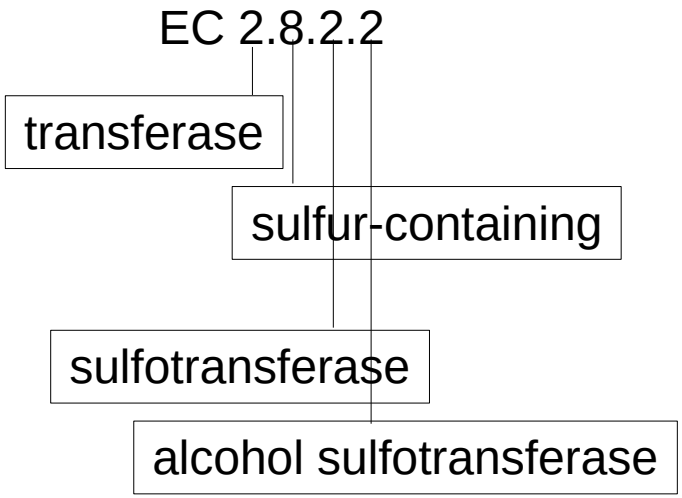
EC 1.1.1.1

main class
1. oxidoreductases
2. transferases
3. hydrolases
4. lyases
5. isomerases
6. ligases
7. translocases

subclass
often tells you often what type of compound or bond it acts on

sub-subclass
further classification

serial number
"just" makes it so each enzyme has its own unique EC number



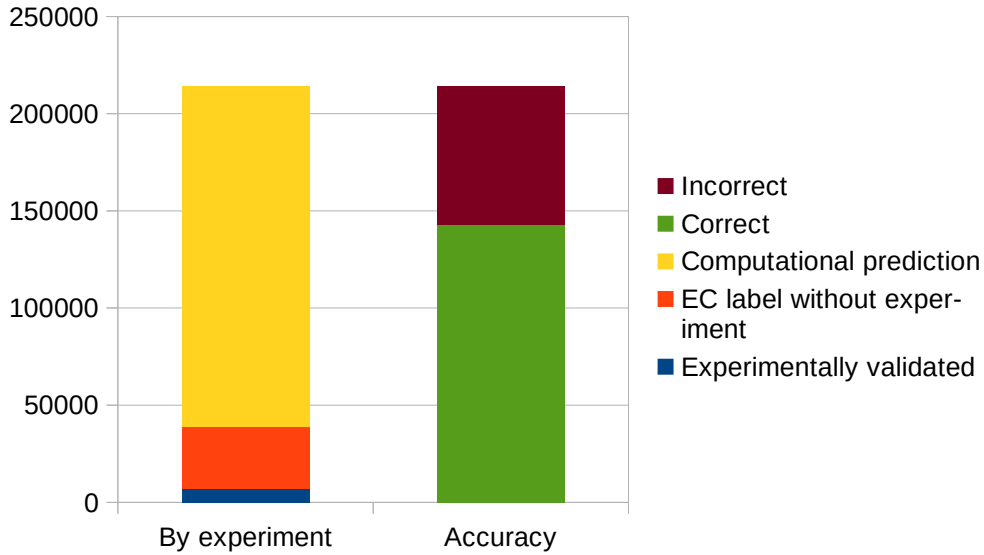
a quick class in **enzyme classes**

1	**oxidoreductases** $AH_2 + B = A + BH_2$	catalyze oxidation & reduction (redox) reactions
2	**transferases** $AX + B = BX + A$	transfer functional groups (such as amino or methyl groups)
3	**hydrolases** $A\text{-}B + H_2O = AH + BOH$	use water to break bonds (catalyze hydrolysis)
4	**lyases** $A{=}B + X\text{-}Y = \underset{X}{A}\text{-}\underset{Y}{B}$	add or remove things to make (or break) bonds (often double bonds, but not always)
5	**isomerases** $A = B$	help molecules rearrange their atoms
6	**ligases** $A + B + NTP = A\text{-}B + NDP + P$ (or NMP + PP)	join 2 molecules with the help of a nucleotide triphosphate (such as ATP)
7	**translocases** $AX + B_{(\text{side 1})} \mid\mid = A + X + \mid\mid B_{(\text{side 2})}$	move things (ions, etc.), often across a membrane

https://thebumblingbiochemist.com/365-days-of-science/enzyme-classes-and-the-ec-numbering-system/

BIIIIIG CAVEAT

- Most functions in databases are not experimentally determined!
  - SwissProt
  - manually curated!



https://academic.oup.com/bioinformatics/article/34/13/i304/5045799

Questions?